USING SOCIAL NETWORKS TO DETECT MALICIOUS BANGLA TEXT CONTENT

BY

NADIM AHMED
ID: 152-15-5869

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
DECEMBER 2018

APPROVAL

This Project titled **"Using Social Network to Detect and Prevent Malicious Bangla Text Content**", submitted by Nadim Ahmed to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 11th December, 2018.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor and Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Ms Subhenur Latif, Assistant Professor, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Ms Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

Co- Supervised by:

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Submitted by:

Nadim Ahmed
ID: -152-15-5869
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes me possible to complete the final year project/internship successfully.

I really grateful and wish our profound our indebtedness to **Ms. Subhenur Latif**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Text Mining*" to carry out this project. Her endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to the Almighty Allah and Head**,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank my entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Social spam has rapidly increased over recent years. Facebook and YouTube contain the most spam content compared with other social media networks. This kind of spam contents like text messaging or comments has a gigantic negative effect on normal user's experience in social media. In this project, I used Naïve Bayes classifier, a supervised machine (SVM) learning algorithm to detect Bangla spam text content. Many spam detection works have been done on English. But I have worked on Bangla language which is used by the most Bangladeshi users. My analysis first collects Bangla text data from YOUTUBE, FACEBOOK and other social media. Then I applied a number of classifiers like Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes etc. At the end, I verified and compared the detectability of Bangla spam text content through different experiment and evaluation. Experiments showed that the Multinomial Naïve Bayes (MNB) algorithm had the best accuracy compared to other machine learning algorithms and my research showed 81.44% accuracy in detecting spam text content from Bangla language.

TABLE OF CONTENTS

CONTENTS **PAGE**

Board of Examiners……i

Declaration…….....ii

Acknowledgements…….....iii

Abstract……...iv

List of Figures…….vii

List of Tables…… ….viii

CHAPTERS

CHAPTER 1: INTRODUCTION……..……..……....1-3

1.1 Introduction…….....1

1.2 Motivation…….....1

1.3 Rationale of the Study……2

1.4 Research Questions……3

1.5 Expected Output…….....3

1.6 Report Layout…….....3

CHAPTER 2: BACKGROUND…….....4-7

2.1 Introduction**……**4

2.2 Related Works…….....4

2.3 Research Summary…….....6

2.5 Scope of the Problem……..6

2.6 Challenges……...7

CHAPTER 3: RESEARCH METHODOLOGY…….....8-15

3.1 Introduction …….....8

3.2 Research Subject and Instrumentation……..…….....8

3.2.1 Research Subject ……...……..8

3.2.2 Instrument …….……10
3.3 Data Collection Procedure…….10
3.4 Methodology and Data Analysis….....11

3.4.1 Pre-Processing..…….....12
3.4.2 Feature Extraction……...…13
3.4.3 Training……..…….....……..13
3.4.4 Algorithm…….. ……14

3.5 Implementation Requirements…….…….……..15
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION..16-19
4.1 Introduction…….....16
4.2 Experimental Results……..16
4.3 Summary…….19
CHAPTER 5: CONCLUSION……..……..……..20-21
5.1 Summary of the Study…….…….……..20
5.2 Conclusion……..……20
5.3 Recommendation……..….....20
5.4 Implication for Further Research…….……...21
5.5 Future Works……..…….21
REFERENCES……...……....22-23

LIST OF FIGURES

FIGURES **PAGE**

Figure 3.1 Raw data…….....11

Figure 3.2 The detail procedures of text classification in terms of a block diagram…...12

Figure 3.3 Naïve Bayes Algorithm (Multinomial Model): Training and Testing……...14

Figure 4.1 Result of confusion matrix ……17

Figure 4.2.1 Pie chart for accuracy rate……...18

LIST OF TABLES

TABLES **PAGE**

Table 3.1 Text category based on interpretation ……9

Table 3.3.1 Bangla Spam Dataset ……10

Table 4.2.1 Confusion Matrix ……..……16

Table 4.2.2 Precision, Recall, F-Score, Error and AUC values…….18

CHAPTER 1
INTRODUCTION

1.1 Introduction

Social Media plays an important role in communications in digital Bangladesh. Social network sites are basically represented by Facebook, Twitter, YouTube, and many others. At present, people spend lots of time on social media. Like Celebrities, public figures, business icons create their social pages for interacting with online users and their fans. But lots of malicious behaviors on the social media makes many troubles to the users.

Social Networks (SNs) have become an important part of users social identity. The initial intent of SNs was to facilitate the connection and sharing. So, People are heavily dependent on online interactions for communications. The increases in content in social media are responsible for the increases of social spams. But unfortunately, this wealth of information, as well as the ease with which one can reach many users and also attracted the interest of malicious parties. Even social networking sites do not provide any strong authentication mechanisms to find out the spammers. Experts estimate that as many as 40% of social network accounts are used for spam [1].

So in the beginning, I along with my team collected lots of samples of spam and ham from real-life uses in social media in order to create the training dataset. Then a detailed filtering process of the naive Bayes classification were explained and I applied this method on my samples for testing at the end of the project. These samples were tested throughout the project using the other methods I will discuss. Although several machine learning algorithms have been employed but probably due to their simplicity and accuracy, I implemented Multinomial Naïve Bayes (MNB) algorithm, accuracy rate 81.44%.

1.2 Motivation

Social networks have lifted the communication system to the utmost level. People spend most of their times Facebook, Twitter, YouTube etc. rather than search engines. In Bangladesh and people who speak Bangla, most of them prefer using Bangla for

networking and developing communication. Content sharing, content contribution, comment or other feedback system is being used to interact with online users by business entities or other public figures. They set up their social pages to enhance direct interaction. However, at the same time, social media networks become susceptible to different types of unwanted and malicious Bangla spam through text contents. Spammers destroy the network environment and this degrades the user's experience of using the network. There is a crucial need in the society and industry for saving the image and maintain a healthy environment in social media. That's why I have decided to work on spam Bangla Text content. I have gone through many spam detecting research papers which had been done in English. Work on Bangla language is very few to detect emotion. So, I thought to work in Bangla spam text. In this demo, I propose a scalable and online social media based Bangla spam content detection system for social network security.

1.3 Rationale of the Study

Due to textual complexity, detecting spam text from Bangla language is very hard. There have been extensive researches conducted for the spam analysis of English texts [18] which showed promising results. This was possible after the advent of the World Wide Web which made a lot of textual data instantly available in electronic media. Before this period, it was hard to develop training data to test theories and models. However, spam analysis of Bangla texts is still a new area and there is a scope of improvement. There are more than 160 million native Bangla speakers and huge amounts of Bangla texts are generated online. Most researches on Bangla texts are performed using news corpus and blogs which are basically extracted by scraping the websites. Another source of data is social media where the opinionated texts are shorter in length but they are informal and full of grammatical and spelling errors and in mixed languages and characters. This Bangla text contains malicious links which misguide the users to fraud and phishing websites. I collected almost two thousand Bangla sentences consisting both positive and negative content. I categorized them into two polarities: spam noted by 1 and ham noted by 0. For example, "মেডাম ছাত্রের সাথে যা করল, দেখুন ভিডিও সহ", "ভিডিও টি একা একা দেখবেন কিন্তু" this two sentences are spam links which redirect the users to a false news or websites. Like these, I have also collected

ham like "ভাল হইতে পয়সা লাগে না", "এরাই দেশের ভবিষ্যৎ" to train the data sets. Then I applied multinomial naïve Bayes classifier to complex pattern recognition and approximation of the function.

1.4 Research Question

Question 1: Does every Bangla sentence have distinct exposition e.g. positive and negative?
Question 2: Does every negative interpretation of the sentence contain spam?
Question 3: Does spam sentence contain some specific word?
Question 4: Can we identify Spam from every new generated informal Bangla sentences?

1.5 Expected Outcome

As there have no work been done to detect malicious Bangla Text content, I have decided to go for it. Our expected output is very satisfactory. First, I will train whether the sentence is spam or ham through MNB algorithm and after the analysis, it will detect whether it is spam or ham.

1.6 Report Layout

The paper is organized into five sections. Following this introduction, Chapter 2 provides brief background details of spam detection field from an information systems perspective, a survey on text analysis those have been published in different information system journals, also the scope of the problem and its challenges. A detailed description of the research methodology including the procedure of data collection, pre-processing and feature extraction is provided in chapter 3. Chapter 4 presents the experimental result of the applied methodology, a brief description of the analysis. And finally, Chapter 5 describes the summary of the empirical research, important limitations of the approach, the implication for further Study.

CHAPTER 2
BACKGROUND

2.1 Introduction

Extensive researches have been conducted for the spam analysis of English texts which showed promising results. It was possible right after the advent of the World Wide Web which made a lot of instant textual data available in electronic media. Before this age, it was very tough to develop training data to test models and theories. However, spam analysis of Bangla texts is still a new area and there is a scope of improvement. There are more than 160 million native Bangla speakers and lots of Bangla texts are generated online. As a result, it would be easier to check the polarity; how much positive or negative the sentence is. After analyzing the text pattern, the sentence could be categorized according to the polarity it belongs to.

In my research, I have mainly researched on how I can detect whether the sentence is spam or ham from a given Bengali text which has been collected from social media such as YouTube, Facebook, FB group like DSU, Murad Takla (মুরাদ টাক্লা) etc. I along with my team collected lots of samples of spam and ham from the real-life uses in social these media in order to create the training dataset. Then a detailed filtering process of the naive Bayes classification would be explained and I applied this method on my samples for testing at the end of the project. These samples will be tested throughout the project using the other methods I will discuss. Although several machine learning algorithms have been employed but probably due to their simplicity and accuracy, I implemented Multinomial Naïve Bayes (MNB) algorithm.

2.2 Related Works

My work is inspired by Chen Liu and Genying Wang's work [2]. In [2], they present an ELM-based spam accounts detection model for social networks. In my work, I am going to implement Multinomial Naive Bayes Classifier. They collect messages crawling from Sina Weibo and then, select three categories of features extracted from message contents, social interactions and user profile properties applied to the ELM-based spam accounts

detection algorithm. In my work, I have categorized our features into 2 properties i.e. spam and ham. I chose Multinomial Naive Bayes classifier to get the optimum outputs.

In [3], Wafa Wali et. Al[3] have proposed a model to measure sentence similarity based on semantic and syntactic-semantic knowledge. Several methods have been proposed to measure the sentence similarity based on syntactic and/or semantic knowledge. Most of the Natural language processing work on sentence or word similarities have been done on English. There are few works which have been done on Bangla and basically, it is done on Bangla blog's[4] and Newspapers[5].

I have chosen the Multinomial Naive Bayes Classification Algorithm because Naive Bayes classifier is very efficient since it is less computationally intensive (in both CPU and memory) and it requires a small amount of training data. Moreover, the training time with Naive Bayes is significantly smaller as opposed to alternative methods [6].

It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, sexually explicit content detection, language detection, and sentiment detection[6]. In[7], Tiago et. Al[7], they proposed and then evaluated a text processing approach for semantic analysis and context detection. They[7] evaluated their approach with a public, real and non-encoded dataset along with several established machine learning methods which can enhance instant messaging and SMS spam filtering.

Naïve Bayes (NB) classifiers is particularly popular among others in commercial and open-source spam filters due to their simplicity that makes them easy to implement, their accuracy and linear computational complexity which is comparable to that of more algorithms in spam filtering[8].

In their papers, Sahami et Al.[9] used a Naïve Bayes classifier with a multi-variate Bernoulli model, a form of NB which relies on Boolean attributes. On the other hand, Pantel and Lin[10] adopted the Multinomial form of NB that normally takes into account term frequencies. It has been shown experimentally in [11] that Multinomial Naïve Bayes performs generally better than the Multivariate Bernoulli NB in text classification.

Vangelis et Al. [12] adopted an experiment which emulates incremental training of personalized spam filters. They [12] made their non-encoded datasets publicly available and are more realistic compared to previous benchmarks. These datasets emulate the varying proportion of ham and spam messages which users receive over time.

2.3 Research Summary

Research is an organized way to find solutions to existing problems or problems that nobody has worked on before. It can be used for solving a new problem or it can be the expansion of past work on any particular field. My research is on detecting spam Bengali text that is associated with NLP(Natural Language Processing).AI(Artificial Intelligence) is challenging the human being to exceed human beings performance. There's been lots of work that has already done to detect spam using texts or documents from various languages. I have studied lots of paper related to detecting spam from a text, lyrics, sentence etc. They used different methods and among them, I have chosen multinomial Naïve Bayes classification algorithm for spam text detection. For that reason, I collected lots of samples of spam and ham from real-life uses in social these media in order to create the training dataset. Then a detailed filtering process of the naive Bayes classification will be explained and I will apply this method on our samples for testing at the end of the project. These samples will be tested throughout the project using the other methods I will discuss. Although several machine learning algorithms have been employed but probably due to their simplicity and accuracy, I implemented Multinomial Naïve Bayes (MNB) algorithm.

2.4 Scope of the Problem

Detecting spam from a text is incipiently a content-based classification which expatiate the concept from Natural language processing (NLP) including Machine Learning(ML) as well. The study of spam detection is very necessary. The increasing number of users in social networks, along with the trust they inherently have in their virtual profile, makes a propitious environment for spammers. In fact, reports clearly indicate that the volume of spam over the social network is dramatically increasing year by year. It

represents a challenging problem for traditional filtering methods nowadays since such messages or links are usually fairly short and normally rife with slangs, idioms, symbols and acronyms that make even tokenization a difficult task. Improved accuracy and consistency in text mining techniques can help to overcome the current problems. Currently, as the next wave of knowledge discovery, text analysis is achieving high commercial values. In this research, I will analyze Bengali text from Facebook status, YouTube comments etc. for finding associated spam of each sentence like positive or negative. After identifying the polarity of each sentence I will then try to find spam text content of each sentence.

2.5 Challenges

Detecting spam or ham from Bangla text content provides huge challenges. some of the sentences like "তারে দেখতে তো ব্যাঙ্গের বাচ্চার মতো দেখায়, সে নাকি আবার হিরো আলম". Here "তারে দেখতে তো ব্যাঙ্গের বাচ্চার মতো দেখায়"- is used as abuse. It indicates negativity of the sentence and the system will automatically detect as spam (1). On the other hand, "সে নাকি আবার হিরো আলম" is just a simple sentence which systems detects as ham (0). Though the whole passage indicates a spam behavior, it is very complex to recognize the pattern of each word and sentences. Misspelling, stop words like ',', ' !', '?', '.', ''. '~', '||', 'I' etc degrades the processing which provides a low accuracy rate. Bangla language having a huge vocabulary, words having different meaning and their various uses makes it more complex for text mining. Informal words like " ওয়াক্কক থুউউউউউউউ", "উফফফফফফ" etc. provides another challenge for modeling Bangla text. Because we are doing our research based on the generated expression of the sentences, it is possible to have the same expression with different polarity.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

This chapter will give an outline of research methods that were carried out to detect spam from a given Bengali Text. It provides information about how data can be processed by applying some certain techniques to sort out spam from them. The instrument that is used to extract the spam from Bengali text from Facebook status and other sources is also described and the procedures that were followed to carry out this data extraction are included. It also provides the methods used to analyze the textual data. Lastly, the implementation and requirements that were followed in the process are also discussed.

3.2 Research Subject and Instrumentation

3.2.1 Research Subject

The main goal of this research is to detect spam from a given Bengali text in order to come up with spam detection associated with it by using Multinomial Naïve Bayes classification algorithm. In case of finding the spam of a sentence, text mining analysis can make it very specific. A set of data is been collected and I categorized them into two sections i.e. Spam and Ham. Spam column contains Spam sentences, words whereas Ham column contains sentences which have a positive meaning. Then we used 80% of these datasets for training and else for testing.

Table 3.1 implies the categorization of spam and ham data which have been collected from various social applications' status, comments etc. It is observed that the polarity of different sentences is generated according to the categorization shown in the table 3.1 below:

Table 3.1 Text category based on interpretation

Spam (1)	Ham(0)
ভিডিওটি একা একা দেখবেন কিন্তু	
এলো অপু বিশ্বাসের গোপন ভিডিও	
কথা গুলা খেত টাইপের	মজা পাইছি
দৌলতদিয়ার কর্মীর গোপন ভিডিও	
ছাগলের তিন নম্বর বাচ্চা	ভালই সেলিব্রেটি হইতেছেন।
গ্রামের যুবতী মেয়েরা দেখুন কি করে । গোপন ভিডিও ফাস	
	শিক্ষিত নয় সুশিক্ষিত হও
	জোর যার মুল্লুক তার
শিক্ষক ও ছাত্রীর গোপন ভিডিও দেখুন	
	বাংলা আমার অহংকার

Whenever a Bangla Sentence is used as input, the system would possibly able to determine whether it is Spam data(1) or Ham data(0) behind the textual content based on the interpretation of sentence pattern. In this experimental study, I have introduced the feature extraction method for detecting malicious Bangla text content.

3.2.2 Instrument

For research purposes, I have collected around 2000 Bengali sentences from different sources like Facebook status, YouTube comments, textbooks, newspaper, direct speech etc. My work is to detect spam from a sentence by applying text classification algorithm. Some well-performed algorithm like ELM, keyword spotting method, support vector machines(SVM), hidden Markov model etc. are used in case of text analysis. Therefore these algorithms give a very high accuracy of almost 90%. In my research, I have used "Multinomial Naïve Bayes" classification algorithm to find the polarity of my test sentences.

3.3 Data Collection Procedure

Even though many datasets in the different language are available in the different databank for research purposes, in terms of Bangla language it is rare. Therefore, I have chosen to build my own dataset from various social media like Facebook, YouTube and named it Bangla Spam Dataset as presented in table 3.3.1

Table 3.3.1 Bangla Spam Dataset

Total Instance	1965
Spam	1319
Ham	646

In order to come up with accurate and objective findings, A good research mainly relied on both primary and secondary data. Primary data's are the raw data which is mainly used

for the original purpose. Those data contained many stop-words like punctuation and special symbols which is directly taken from the field by interviews and questionnaires. I removed those symbols and punctuation to get the secondary datasets. Secondary data is collected for purposes other than the original use. The research has been carried out using secondary data. The main intention was to create a properly trained data set consists of Bengali spam keywords.

Figure 3.1 shows the collection of our raw data which I have collected from different sites like Facebook, YouTube, Newspaper, Blogs etc.

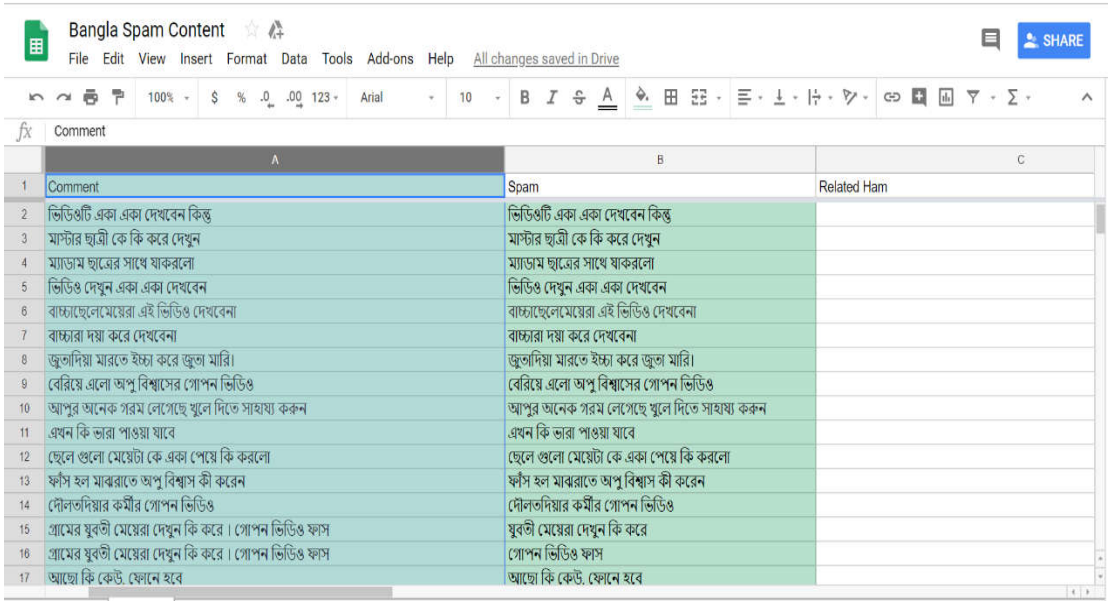

Figure 3.1 Raw data

3.4 Methodology and Data Analysis

Prior to applying categorization techniques to Bangla text with the classifiers, it is inevitable to prepare proper datasets for testing and training. At the same times, pre-processing of Bangla text also required before trainings and construction of model for

successful text categorization. Figure 3.2 illustrates the overall system of Bangla text classification process.

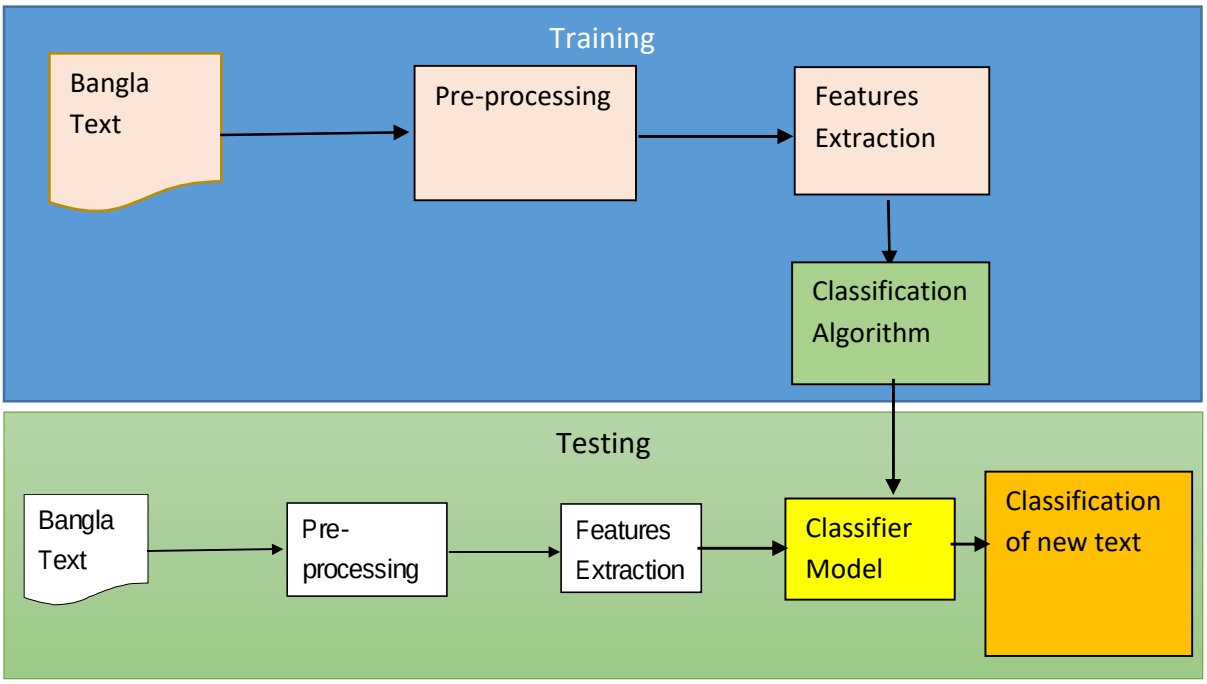

Figure 3.2 The detail procedures of text classification in terms of a block diagram.

3.4.1 Pre-Processing

A proper representation of words within text documents is important to acquire good Classification performance. To Train my model, it required tagged data. I formatted our dataset into two column. One is "Text" that contain actual text data and the other is "Status" that contain value 0 or 1. Spam text is labeled as 1, and non-spam as 0. It's recommended to apply a classification algorithm on cleaned carpus instead of noisy carpus. The noisy corpus includes insignificant things within the text like numerical values, punctuations marks, emoticon etc. Removal of these entities from corpus will increase accuracy because the size of the sample space of possible features set reduced. For example – emoticon like :-P :-D are important while sentiment analysis, but may not be important while others classification. After eliminating all the punctuations marks, numerical value, and emoticon, now we have a clean dataset to fit into a classification algorithm.

3.4.2 Feature Extraction

After the pre-processing phase, I have to extract features from these words prior to applying the algorithm. Different statistical approaches can be used to extract features from this text corpus like Count vectorizer, TFIDF vectorizer etc. Count vectorizer has just counted the frequency of each word. We used TFIDF (term frequency–inverse document frequency) vectorizer to extract the features from the document because it provides a way to score the importance of word based on how frequently they appear across multiple documents.

- If a word appears frequently within a document, give that word higher score.
- If a word appears frequently across multiple documents, that means it's not unique identifiers, give that lower score.

That is how, a common word like "আমি", "তুমি", "ও", "এবং" that frequently appears across many documents will be scaled down, and word that appears frequently within a single document will be scaled up. This will lead towards a better classification performance. The weight for a term i in terms of TF-IDF is given by

$$W_i = \frac{(TF_i \times \log(\frac{N}{n_i}))}{\sqrt{\sum_{i=1}^{n}(TF_i \times \log(\frac{N}{n_i}))^2}}$$

Where N= total number of documents and n_i= document frequency of term i.

3.4.3 Training

With the dataset we got after pre-processing and features extraction, I have to split our dataset into test set, and train set. I split the dataset 80% as train set and 20% as test set. I have to train a classification model. Choosing appropriate algorithm is one of the most crucial point. We choose Naïve Bayes algorithm to train our model. Naïve Bayes is widely used for text classification. When dealing with text, it's very common to treat each unique word as a feature, and since the typical person's vocabulary is numerous thousands of words, this makes for a huge number of features. The simplicity of the algorithm and the independent features assumption of Naive Bayes make it a strong performer for classifying texts. Scikit learn library contain three types of Naïve Bayes Model. Gaussian Naïve Bayes

, Bernoulli Naïve Bayes and Multinomial Naïve Bayes. Which variant of Naïve Bayes should applied, depends on data. Multinomial naive Bayes treats features as event probabilities. It has been shown experimentally in [11] that Multinomial Naïve Bayes performs generally better than the Multivariate Bernoulli NB in text classification. Multinomial NB surprisingly performs even better if term frequencies can be replaced by Boolean attributes [16].

3.4.4 Algorithm

Naive Bayes Classifier works based on Bayesian theorem. Multinomial and Bernoulli distributions are popular while classifying document classification including Spam Filtering. In my case, Multinomial NB do better than Bernoulli. The Multinomial NB work as follows:

TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c/N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

APPLYMULTINOMIALNB($\mathbb{C}, V, prior, condprob, d$)
1 $W \leftarrow$ EXTRACTTOKENSFROMDOC(V, d)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in W$
5 **do** $score[c] += \log condprob[t][c]$
6 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

Figure 3.3: Naive Bayes Algorithm (Multinomial Model): Training and Testing [17].

3.5 Implementation Requirement

We have used Python language for implementation where the platform is Anaconda. The tools are listed following:

i. Anaconda.
ii. Python.
iii. MS Excel.
iv. Notepad++.
v. Socialfy. (Facebook Comment Extractor tools)
vi. ytcomments. (YouTube Comment Extractor tools)

For input insertion, we used Avro keyboard.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This is an experimental based research that I have worked out. In this chapter, the results of malicious spam text content from the Bangla language are presented according to their polarity. In total, I have collected 1965 sentences from Facebook statuses, YouTube comments, Bengali blogs, newspapers, and textbooks. The experiment has been carried by using Naïve Bayes (NB) which includes Pre-processing, feature extraction and finally text classification methods. According to their pattern polarity, spam, and ham, the sentences were identified and the results have been discussed which includes the total accuracy of our experiment in details. After evaluating the polarity results we have finally come out with a satisfactory outcome.

4.2 Experimental Results

After training, using training dataset, it's time to taste out dataset using taste set that is unknown to our model. From results, multinomial naive Bayes yields an overall accuracy of 81.44%. The confusion matrix is shown as follows in Table 4.2.1.

Table 4.2.1 Confusion matrix

	Predicted Non Spam	**Predicted Spam**	**Total**
Actual Non-Spam	194	11	205
Actual Spam	44	47	91
Total	238	58	296

Test results after performing the tests has been pictured in the following Figure 4.1

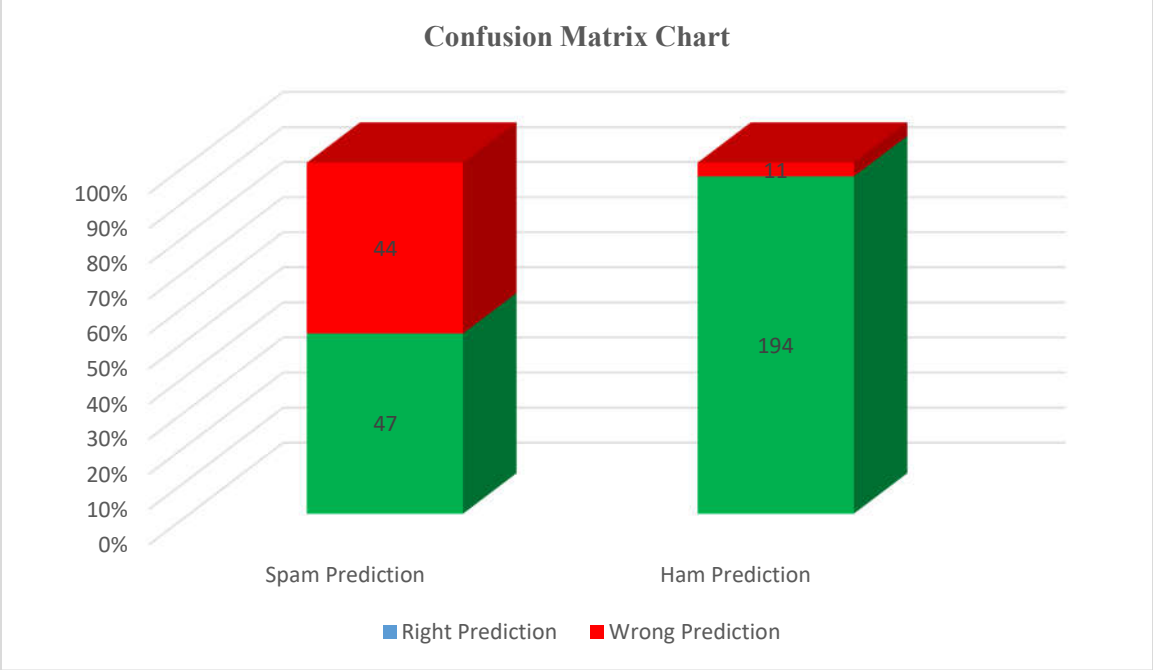

Figure 4.1 Result of confusion matrix

To finalize the total accuracy of our experiment, I have sorted out individual accuracy for test1, test2, test3 and test4. After getting their individual accuracy outcome we then executed the total accuracy which is 81.44%.

Accuracy= (TP+TN)/(TP+TN+FP+FN)

Here TP= True Positive (case was positive and predicted positive)
TN=True Negative (case was negative and predicted negative)
FP=False Positive (case was positive but predicted negative)
FN=False Negative (case was negative but predicted positive)

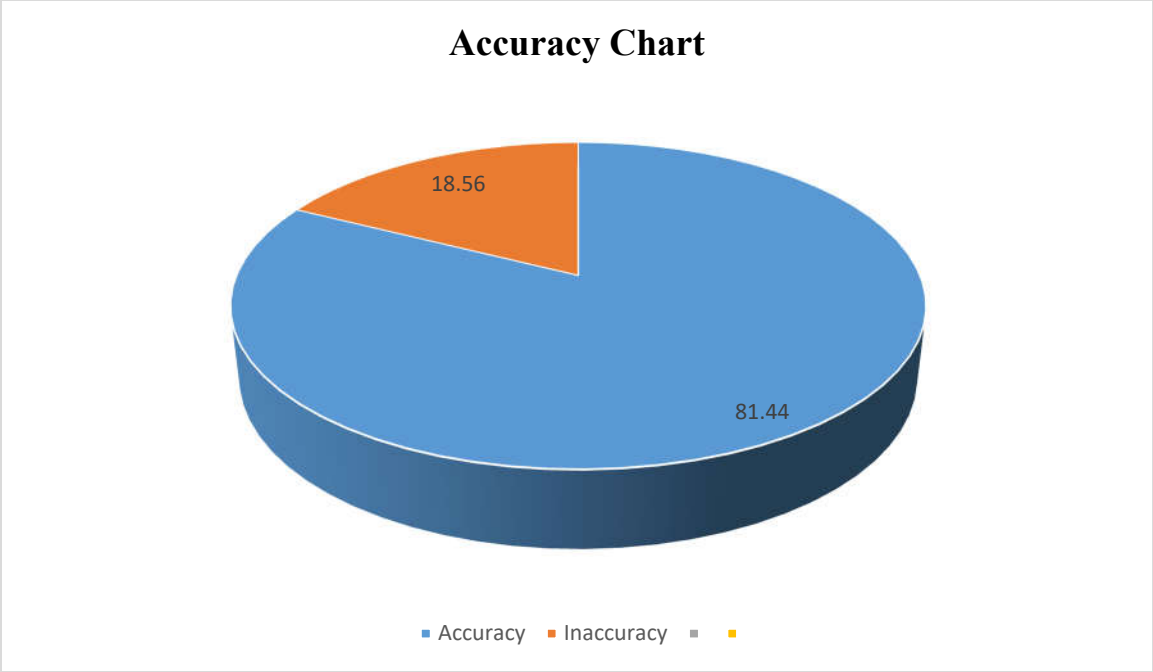

Figure 4.2.1 Pie chart for accuracy rate.

Error, Precision, Recall, F1 Score, AUC (Area Under Curve) are shown in the following Table 4.2.2:

Table 4.2.2 Precision, Recall, F-Score, Error and AUC values

Precision	**0.814**
Recall	**0.814**
F-Score	**0.814**
Error	**17.56**
AUC(Area Under Curve)	**0.73**

4.3 Summary

During the implementation of our system, I noticed that the bigger the number of sentences, the higher are the recall and precision. Therefore, I believe that the enrichment of our database of Bangla sentences can significantly enhance the results. After experimenting, I have found that a sentence may have spam, or it may be ham. Featuring the extraction and text classification, I have used the “Multinomial Naïve Bayes” Algorithm and after the experimental result, I have come up with 81.44% accuracy.

CHAPTER 5

CONCLUSION

5.1 Summary of the Study

During the last few decades' text classification has received an incredible attention from people because it helps to classify spam data and threats. Hence, a lot of work is being done in this domain to find the finest classifier for text classification. From the acquired results in contrast with the pre-processing technique, it is clear that the framework with Multinomial Naive Bayes algorithm performs better as compared to other classifiers. After extensive pre-processing MNB was applied and it comes out to be 81.44 % effective in classifying malicious Bangla Text content.

5.2 Conclusion

Detecting spam from a Bengali sentence was not that easy as people disagree on identifying exact interpretation of the same sentence. My text classification method helps us to detect exact expression that majority people think about. Among different approaches, I have used multinomial naïve Bayes classification algorithm to extract semantic information from a sentence for detecting spam from Bangla text content. Finally, the accuracy came 81.44%.

5.3 Recommendation

In this thesis, I have worked with around two thousand sentences. So, my corpus doesn't have sufficient lexicons. As every day, new data are generating through social media, a collection of new pattern sentences are necessary. So before going for test add necessary keywords to the database. While giving input keep a focus on the spelling of the lexicons and also the removal of digit, punctuation, special symbols and stemming is very important to get the best accuracy. In the case of a spelling mistake, the program will fail to detect spam accurately. So the user may get lower accuracy.

5.4 Implications for Further Research

The demand for data mining analyst is highly appreciated in this modern age. This is because of the presence of abundant amount of data in our surroundings. To be more accurate, it is high time to work with these sorts of complex data, so that a new pattern can be introduced to resolve several critical problems. Spam analysis is one of the fundamental branches of data mining. The experimental study which I have carried out on malicious text detection with a satisfactory outcome is leaving a strong footprint behind my work. It has been observed that works on spam detection in Bangla has a lot of valuable impact in our day to day life. We are living in the 3rd world's modern age. In this modernized world, people are seen very active in social media like Facebook, YouTube etc. Business entities set up their public pages on social networks and enhance their direct interaction with their customers through content sharing, commenting, or through any other feedback system. Celebrities, online sellers or institutional organization also publish their content for direct interaction. It is unfortunate that some spammers spoil the environment by posting unethical staffs or posting abusive comments in their posts which destroys the images. So it is very urgent to stay safe from malicious trap. Prevention of this kind of spam needs to be executed as soon as possible. As there has been no work done for Bangla languages, so my research will bring a revolutionary changes in the field of data science and to Bangladeshi people's perspective. I will further research for detecting and fighting against the spam accounts through this process.

5.5 Future Work

i. Achieving higher accuracy by using classifiers in combination
ii. Developing a technique that can catch the sentimental phrases and train methodology for those spams.
iii. Multilingual spam email classification
iv. Enriching corpus with more words.
v. Add a stemmer to reduce the size of our corpus and improve model performance.
vi. Detecting and fighting spam accounts.

REFERENCES

[1] Go.proofpoint.com, 2018. [Online]. Available: https://go.proofpoint.com/nexgate-social-media-spam-research-report. [Accessed: 10- Sep- 2018]

[2] Chen Liu and Genying Wang, "Analysis and detection of spam accounts in social networks", 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 2016.

[3] W. Wali, B. Gargouri and A. Ben Hamadou, "Enhancing the sentence similarity measure by semantic and syntactico-semantic knowledge", Vietnam Journal of Computer Science, vol. 4, no. 1, pp. 51-60, 2016.

[4] Mohammad Samman Hossain, Israt Jahan Jui and Afia Zahin Suzana, "Sentiment Analysis for Bengali Newspaper Headlines", BRAC University, Dhaka, Bangladesh.

[5] Md. Mahfuzur Rahaman and M. A. Alim Mukul, "Trending News Analysis from Online Bangla Newspapes", Shahjalal University of Science & Technology.

[6] Machine Learning Blog & Software Development News. (n.d.). Retrieved from http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/

[7] Almeida, T. A., Silva, T. P., Santos, I., & Hidalgo, J. M. (2016). Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering.

[8] I. Androutsopoulos, G. Paliouras, and E. Michelakis. Learning to filter unsolicited commercial e-mail. technical report 2004/2, NCSR "Demokritos", 2004.

[9] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization – Papers from the AAAI Workshop*, pages 55–62, Madison, Wisconsin, 1998.

[10] P. Pantel and D. Lin. SpamCop: a spam classification and organization program. In *Learning for Text Categorization – Papers from the AAAI Workshop*, pages 95–98, Madison, Wisconsin, 1998.

[11] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI'98 Workshop on Learning for Text Categorization*, pages 41–48, Madison, Wisconsin, 1998

[12] Jin, Z. (2008). Spam message self-adaptive filtering system based on Naive Bayes and support vector machine. *Journal of Computer Applications,28*(3), 714-718. doi:10.3724/sp.j.1087.2008.00714

[13] Wei, Q. (2018). Understanding of the naive Bayes classifier in spam filtering. doi:10.1063/1.5038979

[14] Issac, B. (n.d.). Spam Detection Approaches with Case Study Implementation on Spam Corpora. *Cases on ICT Utilization, Practice and Solutions*. doi:10.4018/9781609600150.ch012

[15] Schneider, K. (2005). Techniques for Improving the Performance of Naive Bayes for Text Classification. *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science,*682-693. doi:10.1007/978-3-540-30586-6_76

[16] K.-M. Schneider. On word frequency information and negative evidence in Naive Bayes text classification. In *4th International Conference on Advances in Natural Language Processing*, pages 474–485, Alicante, Spain, 2004.

[17] D. P. Bhukya and S. Ramachandram, "Decision Tree Induction: An Approach for Data Classification Using AVL-Tree," International Journal of Computer and Electrical Engineering, pp. 660–665, 2010.